

# Высокопроизводительный промышленный сервис для компьютерного зрения на Python

Алексееенко Григорий, Layer SberDevices



**HighLoad++**  
2022



# Немного о себе



- 1 Senior Data Scientist в Layer с 2021
- 2 Работаю в компьютерном зрении пятый год
- 3 Пишу на Python и C++
- 4 Занимался различными проектами от видеоаналитики для ритейла и умных городов до подсчета свиней на ферме



# Монетизация фото- и видеоконтента: In-Image





## Особенности

- 1 Каждый день новые статьи
- 2 Большое количество просмотров, нужно чтобы при просмотре данные были актуальные
- 3 Декодируются jpeg, png и прочие форматы изображений

# Монетизация фото- и видеоконтента: Visual Search

## Куртка кожаная женская XS



Состав: искусственная кожа  
Код товара: 100012302909

Товар закончился

### Похожие товары в наличии



4 999 руб



11 000 руб

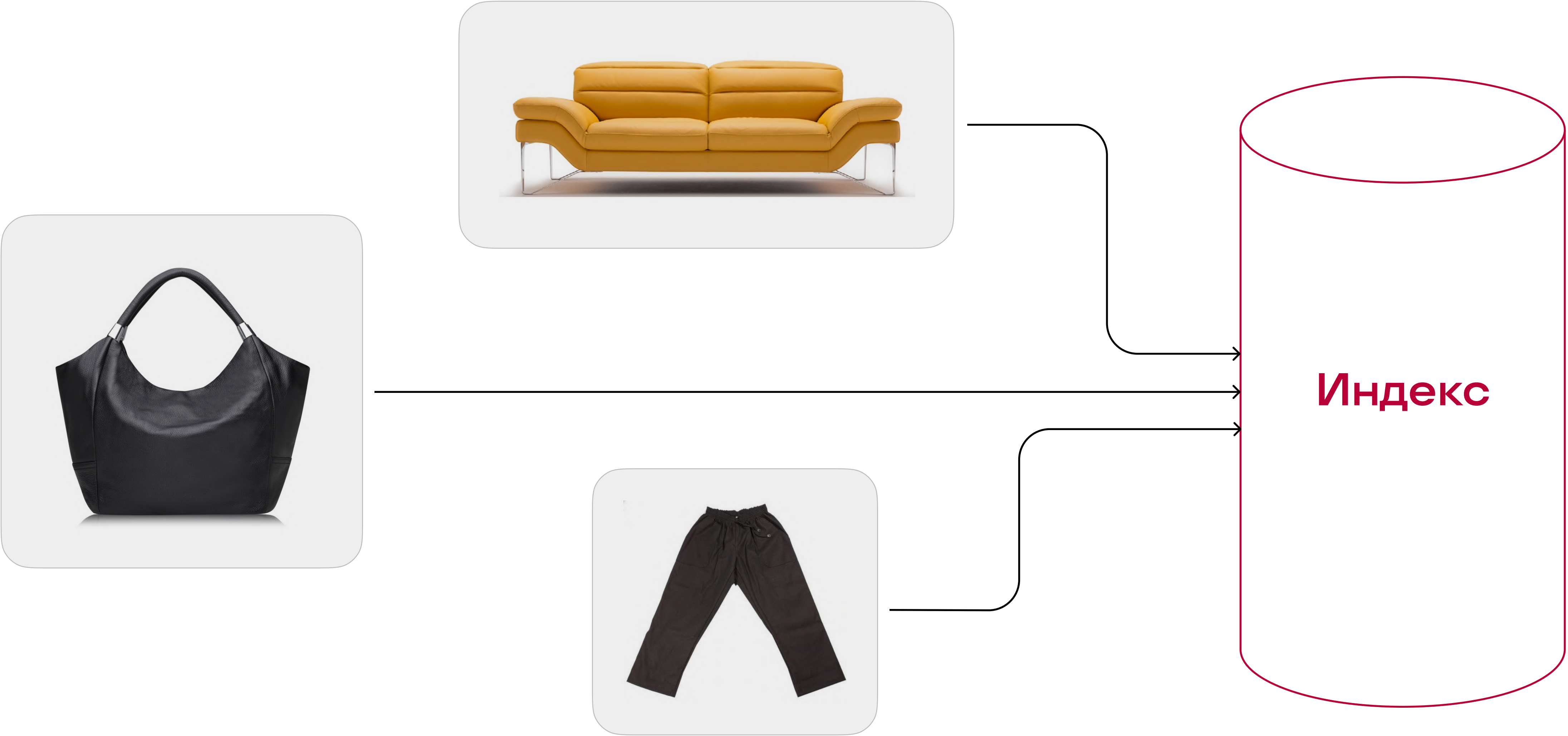


3 000 руб

## Особенности

- 1 Миллионы товаров
- 2 Нужно постоянно обновлять
- 3 Непредсказуемые форматы изображений
- 4 Поиск по большому индексу

# Монетизация фото- и видеоконтента: генерация индекса





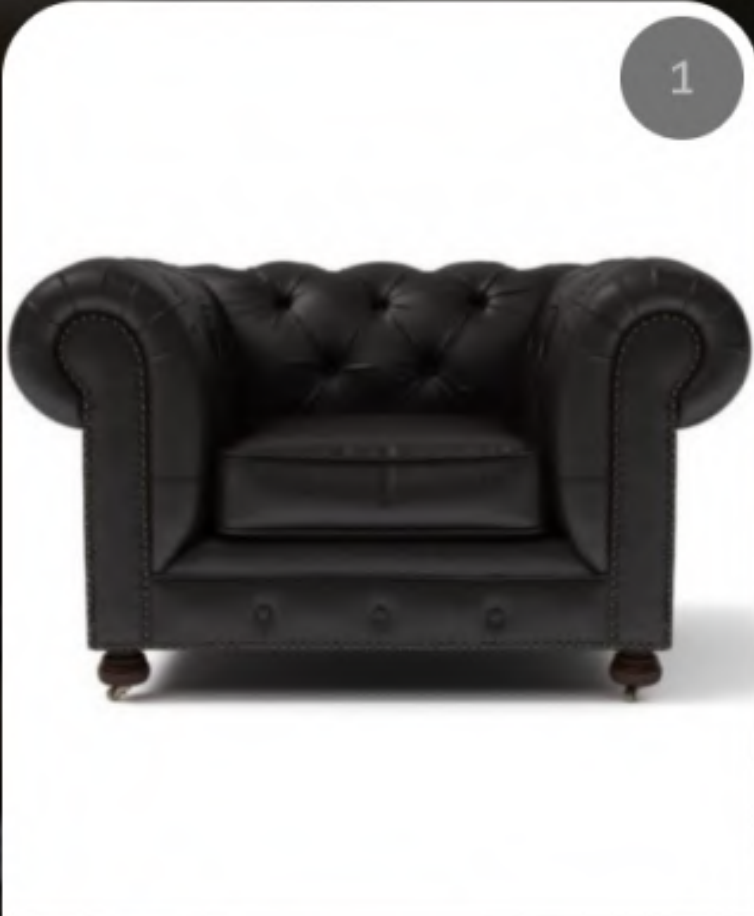
# Монетизация фото- и видеоконтента: In-Video

Актеры

Мебель

Мерч

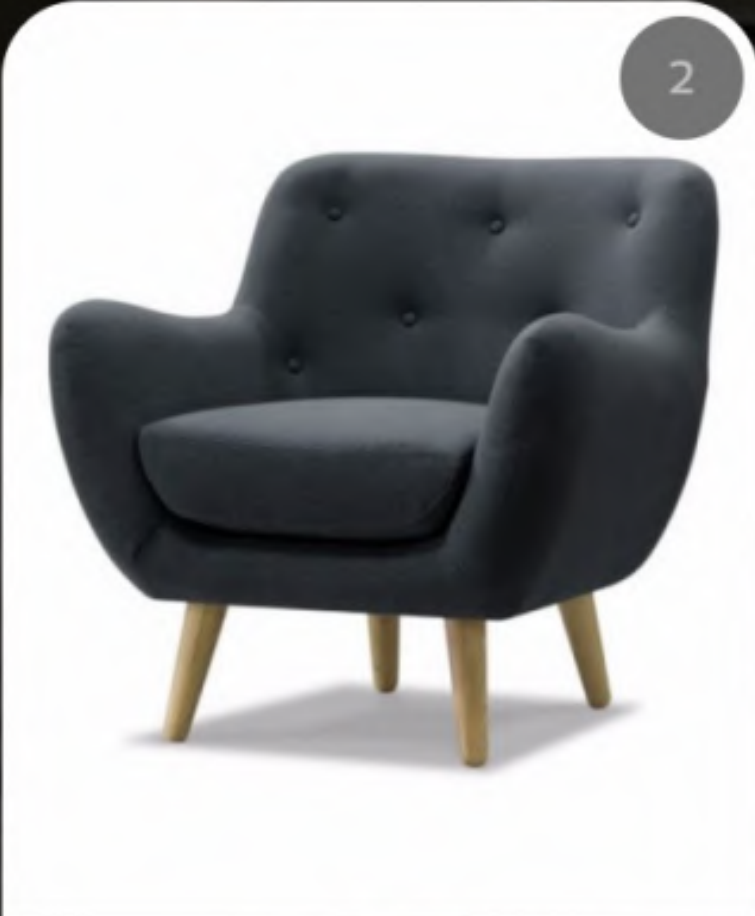
1



Кожаное кресло  
Chester черное

49 000 ₽ ~~59 000 ₽~~


2



Дизайнерское  
кресло Oloff черное

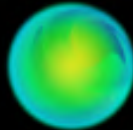
25 500 ₽ ~~31 500 ₽~~


3



Кожаное кресло  
Chester черное


200 000 ₽ ~~210 000 ₽~~







## Особенности

- 1 Работа в реальном времени
- 2 Максимум 40 мс на ответ 
- 3 Нужно декодировать видео

# Нагрузки

## Visual Search Генерация индексов

### Платье женское Viserdi 10047-грф 4111200 серое 52 R

Артикул 100 030 032 899

Размер RU: 44 46 48 50 **52**

**Готовимся к школе!**

**Таблица размеров**

Размер RU: 52  
Размер производителя: 52 RU  
Материал: полиэстер; вискоза; лайкра  
Материал подкладки: без подкладки  
Цвет: серый  
Стиль: офисный  
Особенности: с разрезом спереди  
Код товара: 100030032899

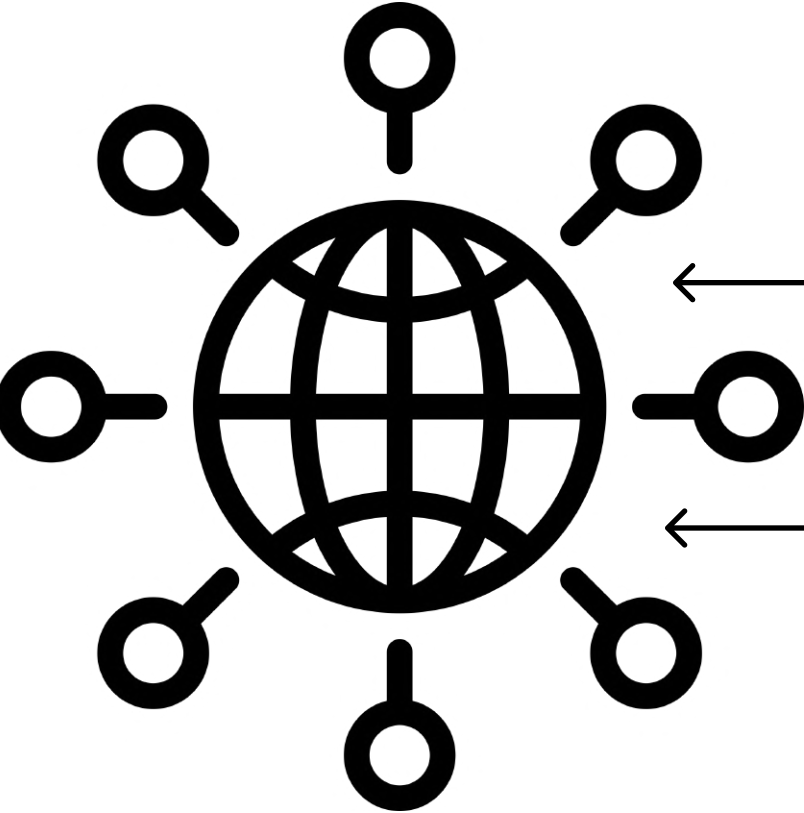
[Все характеристики](#)

**Примерка**  
Покупайте только то, что подошло

**Женские платья**  
**Женские платья Viserdi**  
**Viserdi**

**Похожие товары**

jpeg, webp,...



## In-Image

Жюли Ферри для жаркого лета в Асечо выбирает классические образы в светлых оттенках. Сетчатая для прогулки французской инфлюенсер сделала ставку на светлый жакет в полоску, короткое платье из денима и простую белую майку. Образ дополнила дипломная сумочка Bora карамельного оттенка, винтажный ремешок, солнцезащитными очками Bottega Veneta и золотыми украшениями.

Сексизм Меллер (a Self-Portrait)

Ждет многообещающий гонимый социальными Сабуров объявил о новом... 6 232

Тайна семьи Содар: патро датель Босхиде пропала из дома во время пожара, но... 31 140

Четыре обязательных вещи, которые нужно сделать сейчас, чтобы избежать мифа... 5 172

В 2019 году стриминговая платформа Netflix подписала контракт на производство сериала «Песочный человек» по одноименному комиксу Нила Геймана. Главный герой произведения – повелитель царства снов Морфей, который попадает в плен к колдуну Родригу Берджесу, однако позже ему удается вырваться на свободу. Съемки «Песочного человека» начались в октябре 2020-го – несколько позже, чем планировалось изначально. В планы создателей вмешалась пандемия коронавируса.

**СЕГОДНЯ ЧИТАЮТ**

Имена деятелей, которые были популярны в СССР – есть ли у них шанс вернуться в моду?

В 40 лет жизнь только начинается: главные свадебные платья в жизни Бэры Боли

Обсуждающий Анатолий Белый вновь обратился к россиянам

Тайна семьи Содар: патро датель Босхиде пропала из дома во время пожара, но полиция отказалась их искать

В зоне риска вакцинированные J&J и не только: медком, поразивший Чубакку, впервые пришел к смерти

jpeg

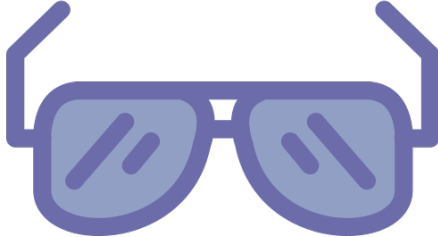
## In-Video

HLS, WebRTC

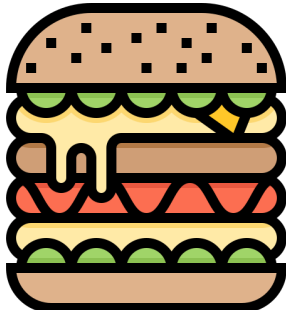





# Layer




Accessories




Food




Locations



Furniture

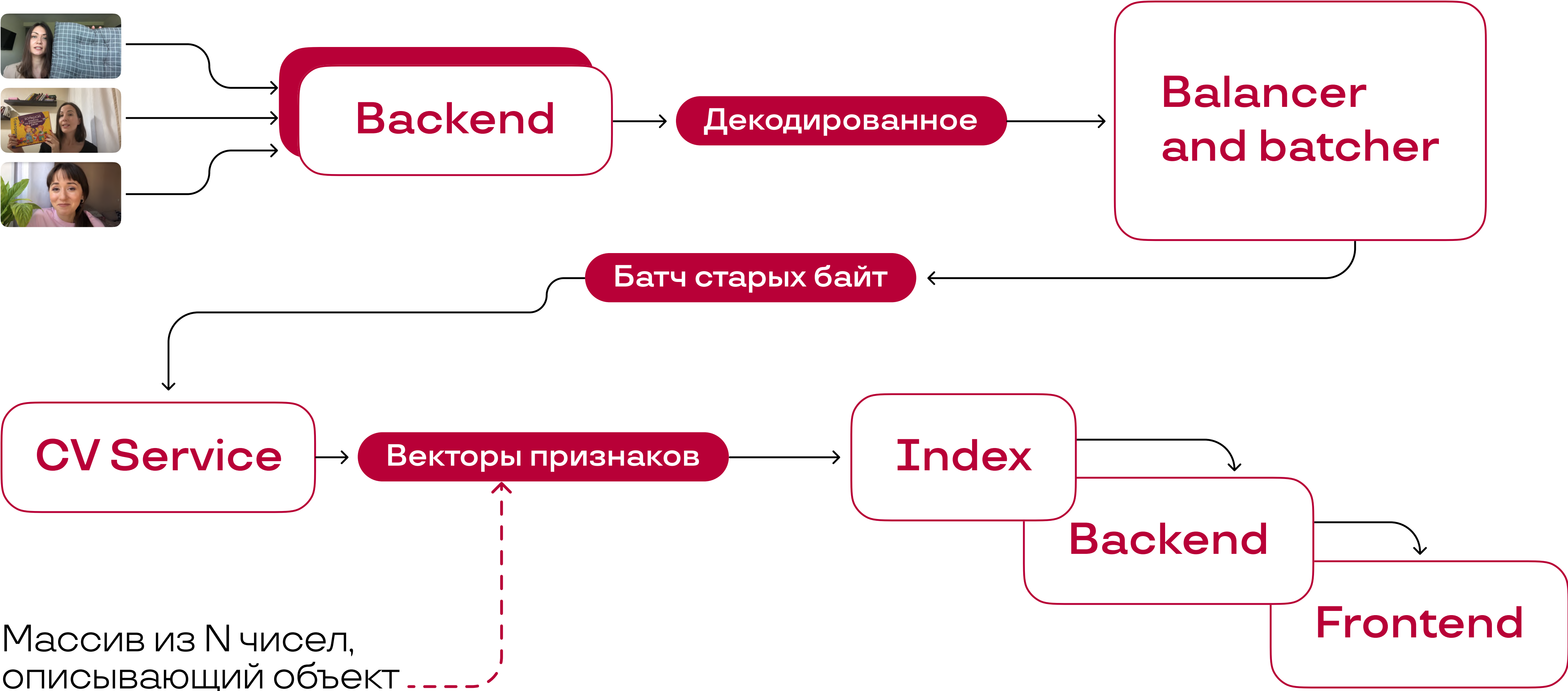


Clothes



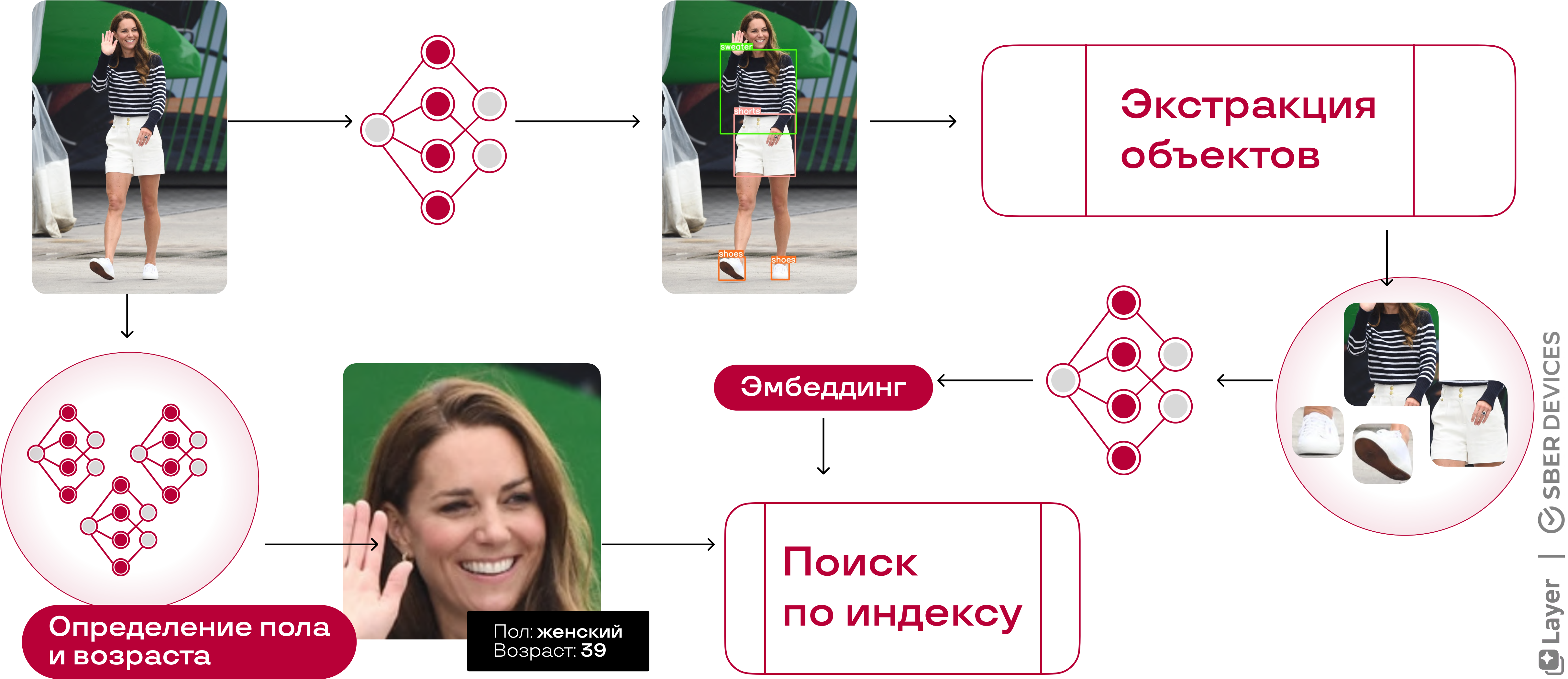
Actors

# Клиент-серверная архитектура





# Детали пайплайна распознавания одежды



# Варианты оптимизации

12

**Модель**

**Инференс и батчинг**

**Вспомогательный код**

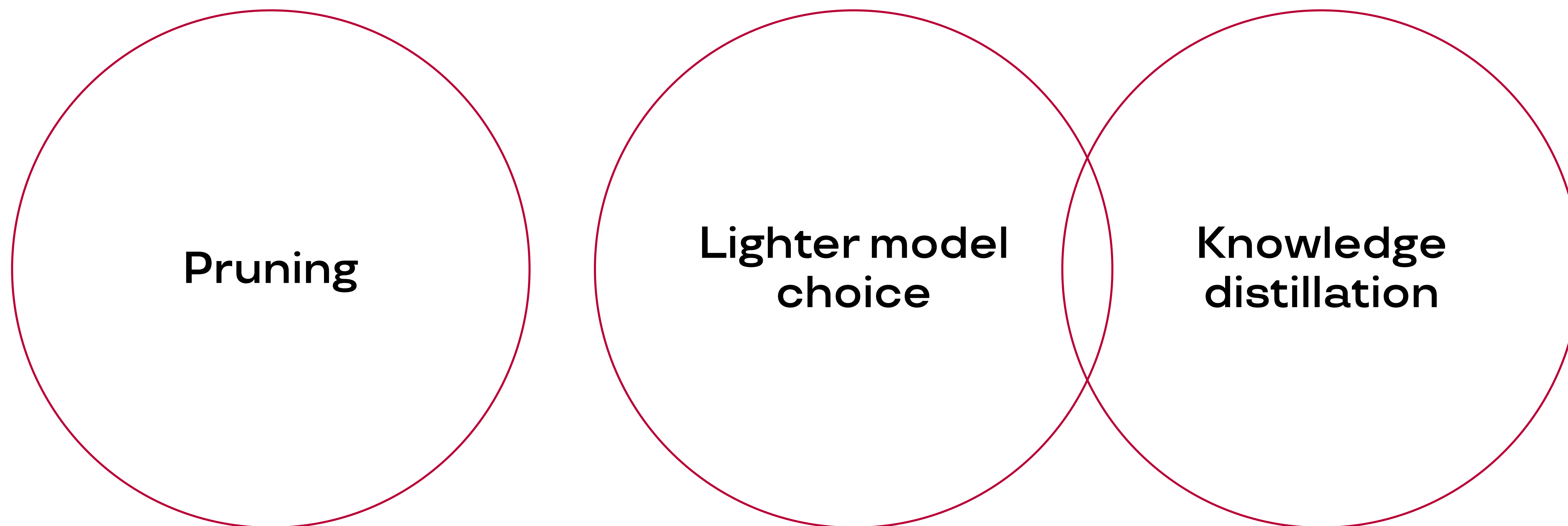


# Варианты оптимизации

**Модель**

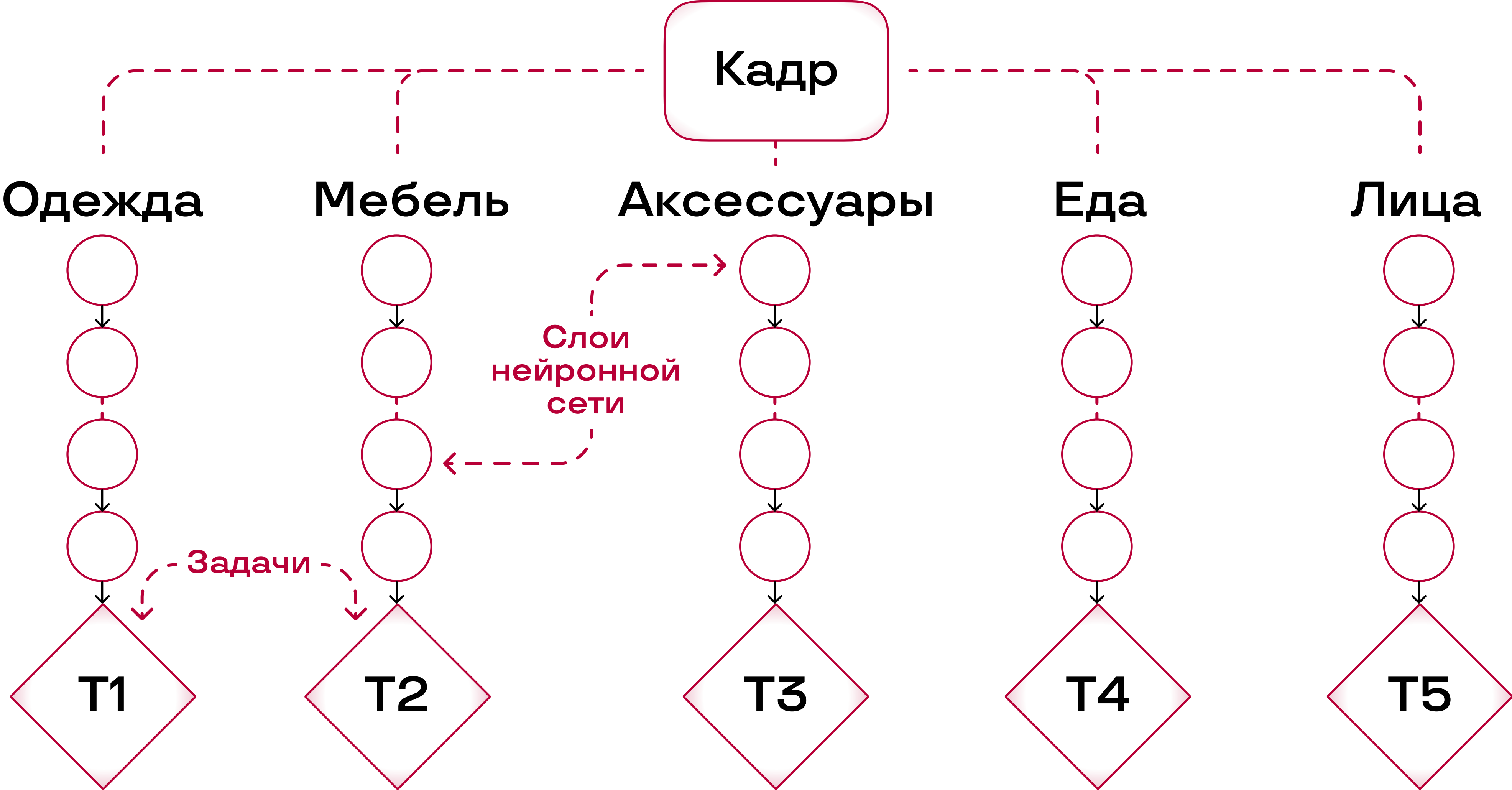
**Инференс и батчинг**

**Вспомогательный код**

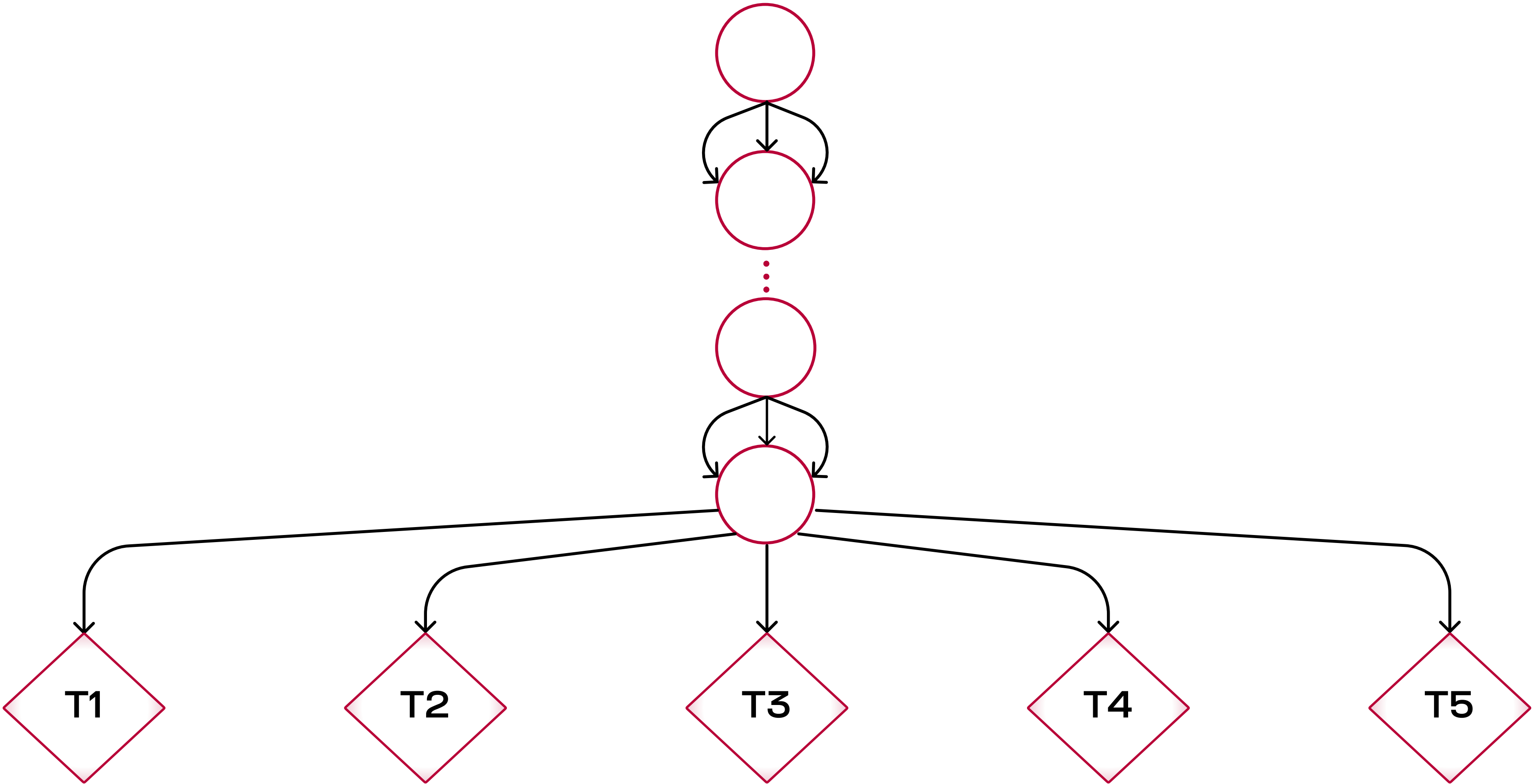




# Оптимизация моделей: многозадачность

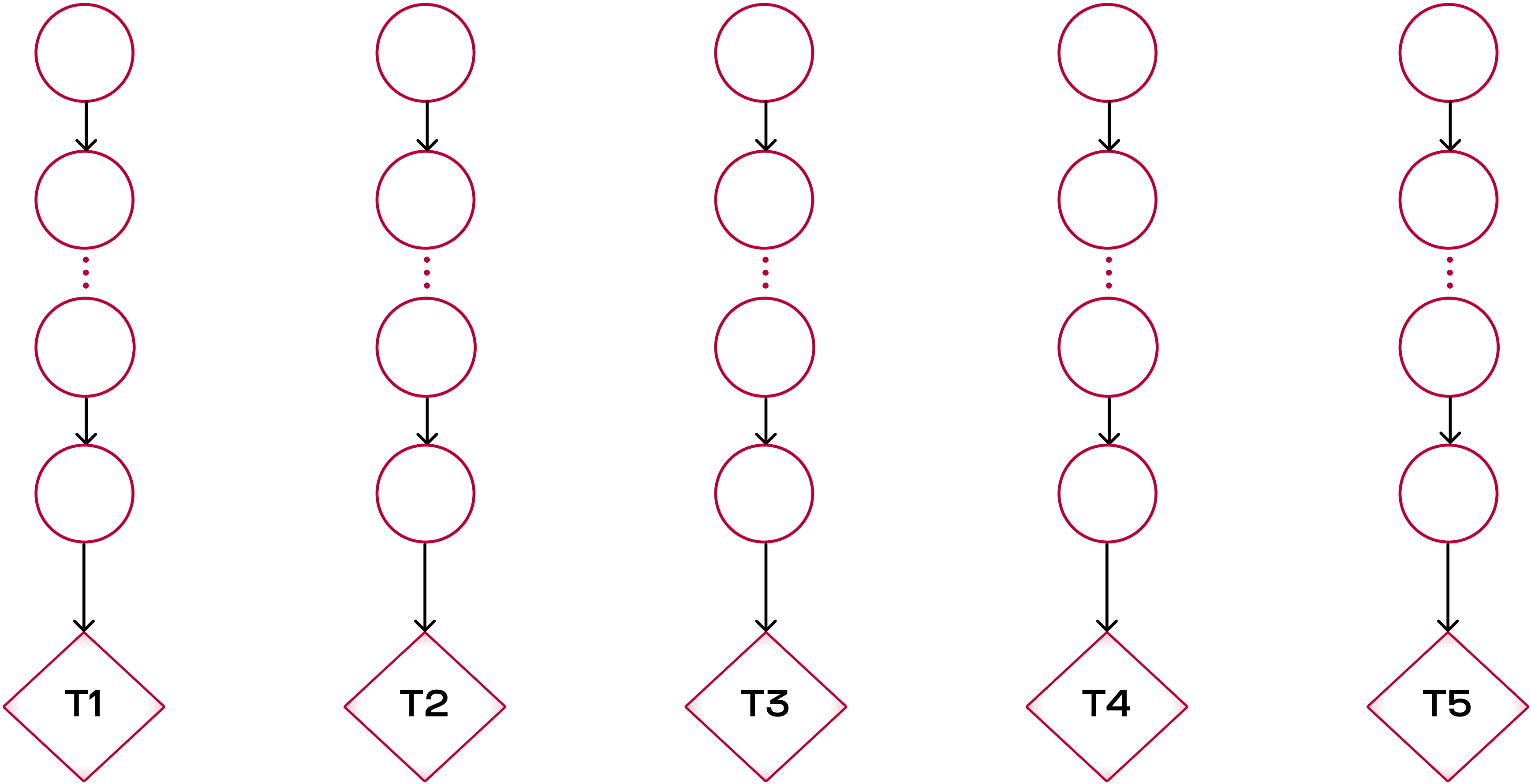


Подход «в лоб»

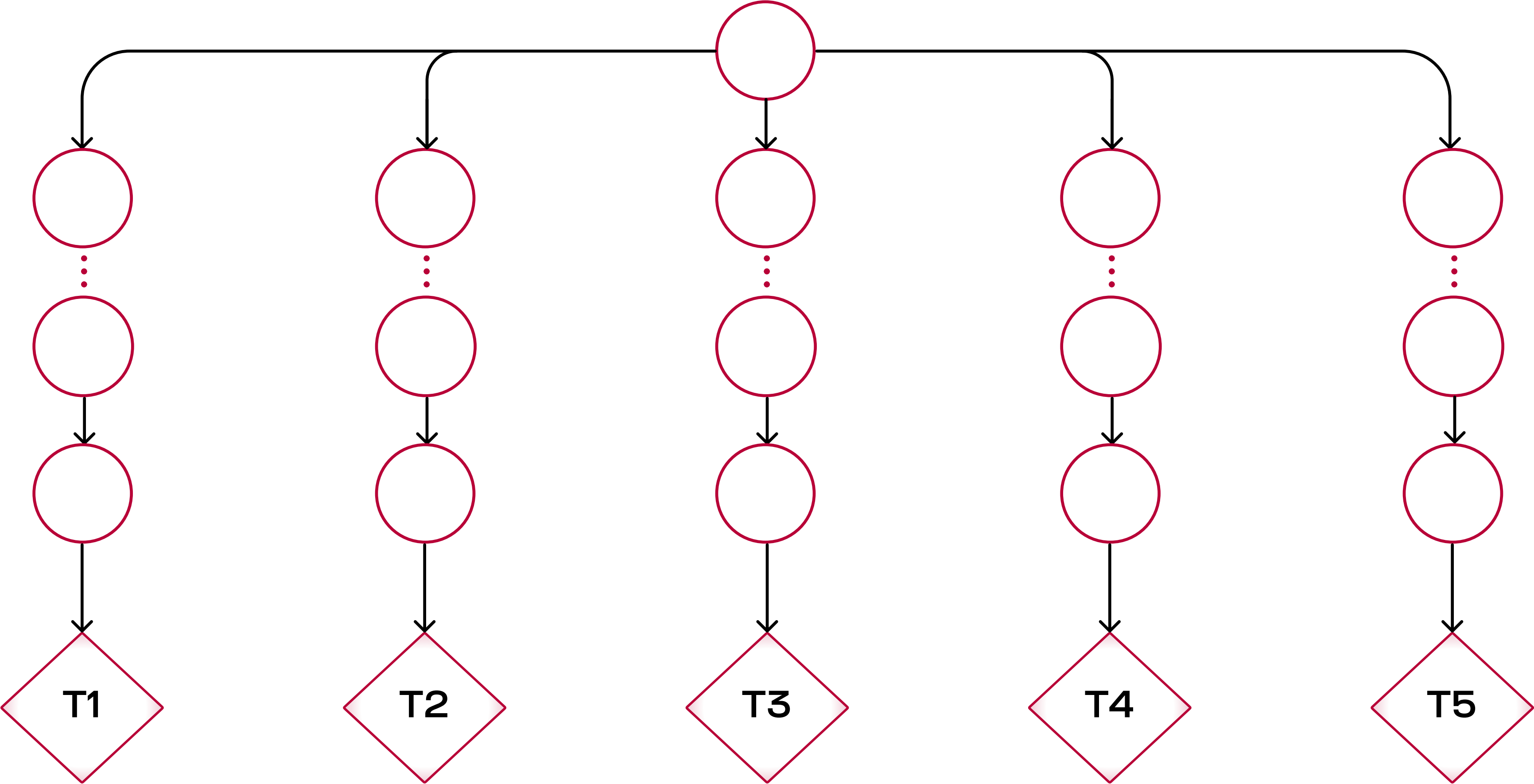




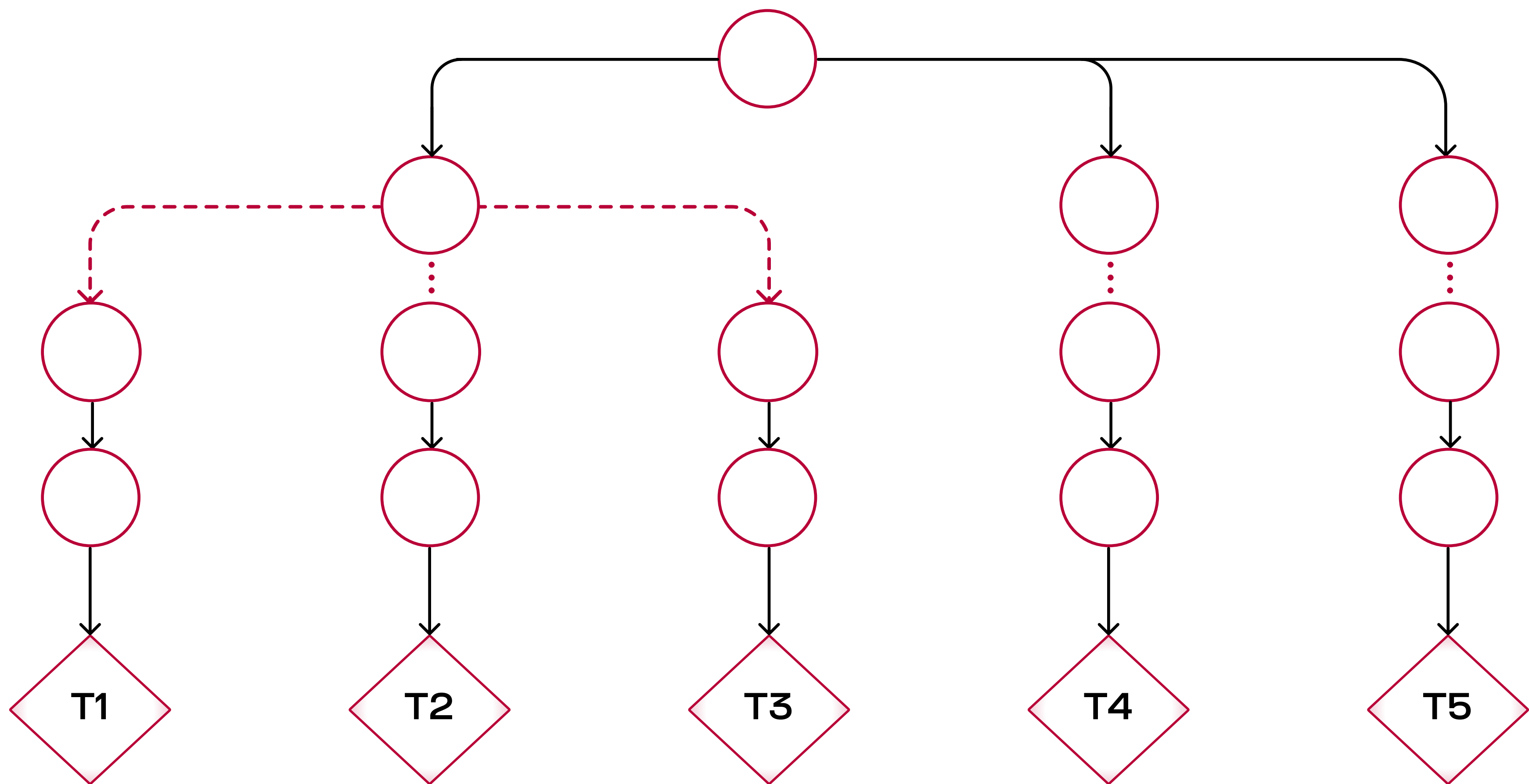
# Оптимизация моделей: многозадачность



# Оптимизация моделей: многозадачность

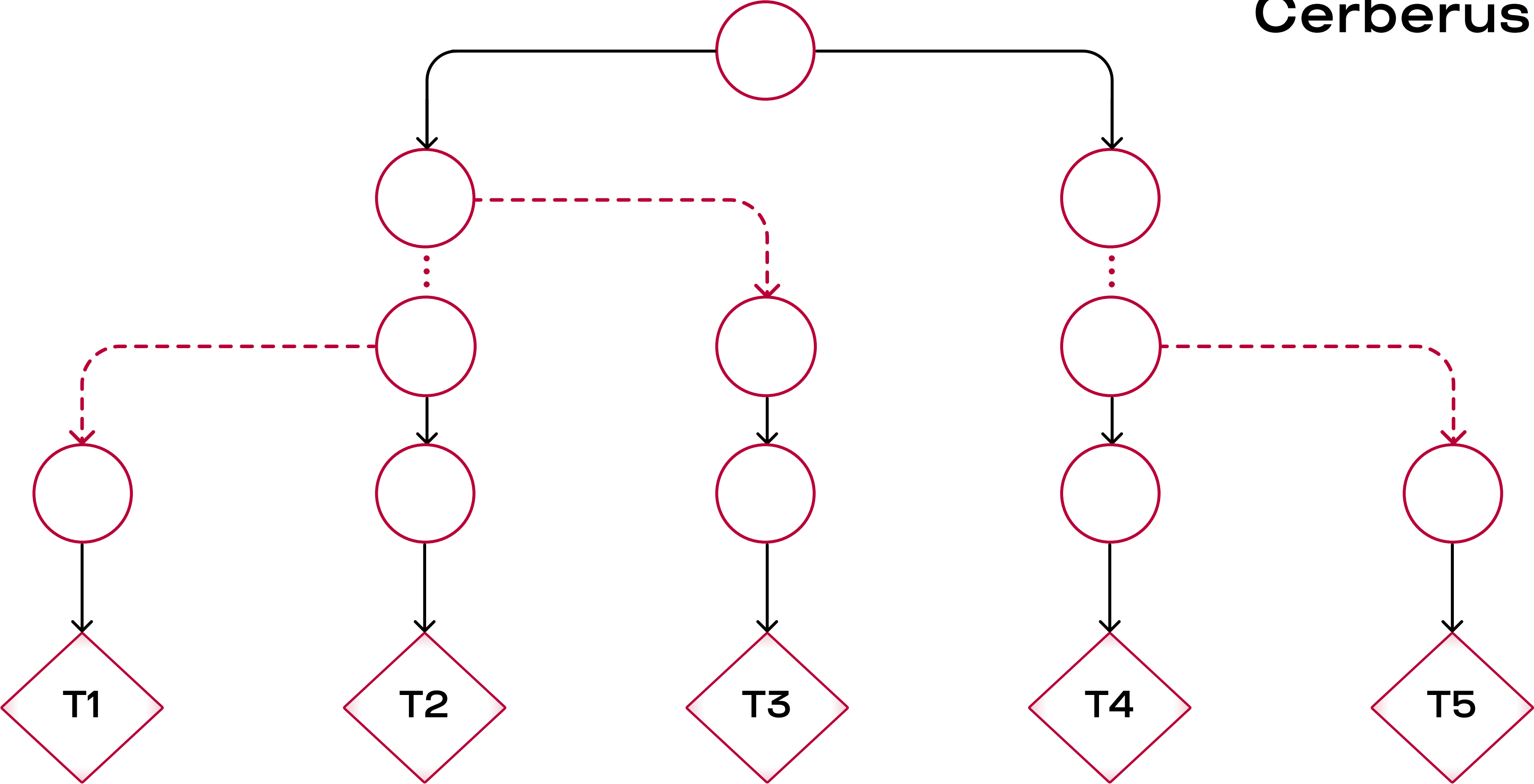


# Оптимизация моделей: многозадачность





Cerberus Net



# Варианты оптимизации

21

**Модель**

**Инференс и батчинг**

**Вспомогательный код**

# Варианты оптимизации

22

Модель

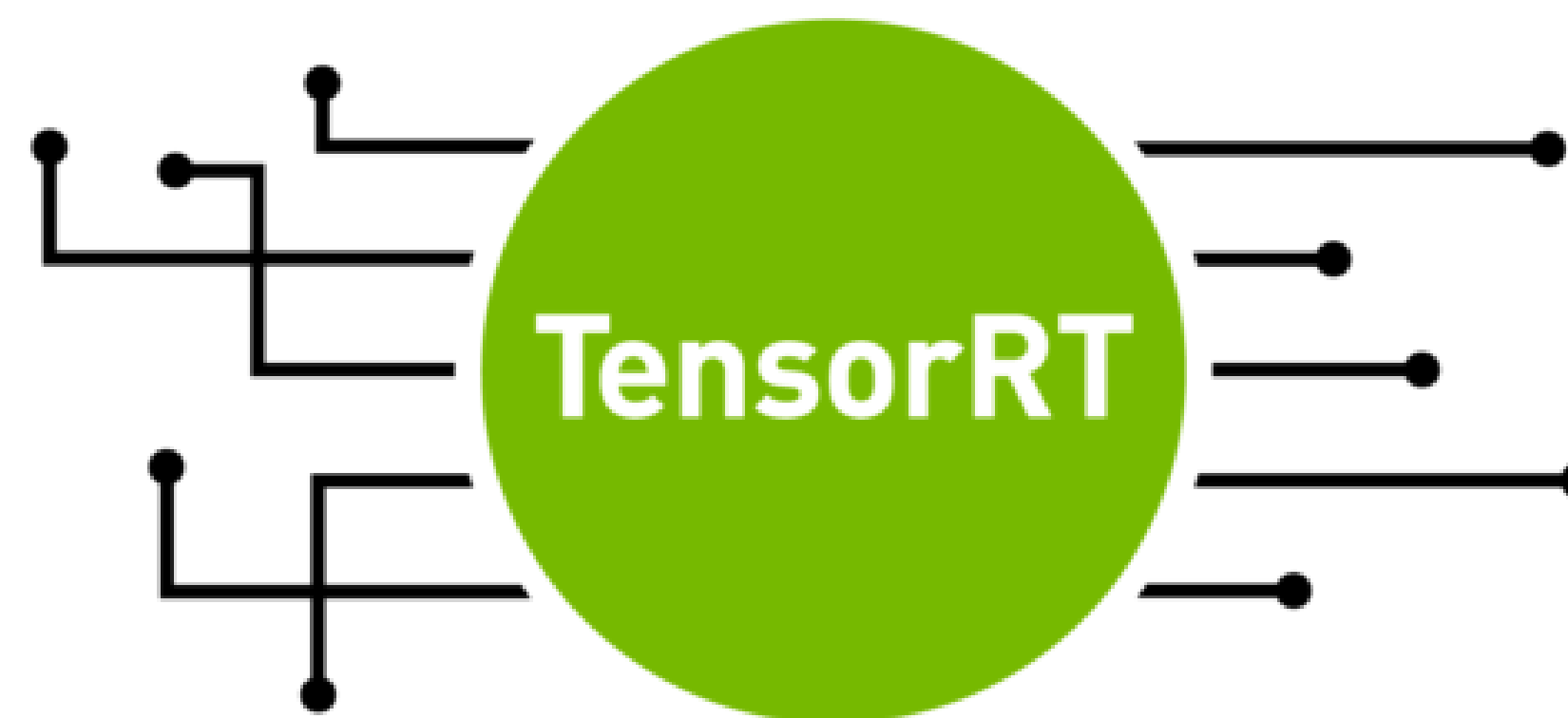
Инференс и батчинг

Вспомогательный код

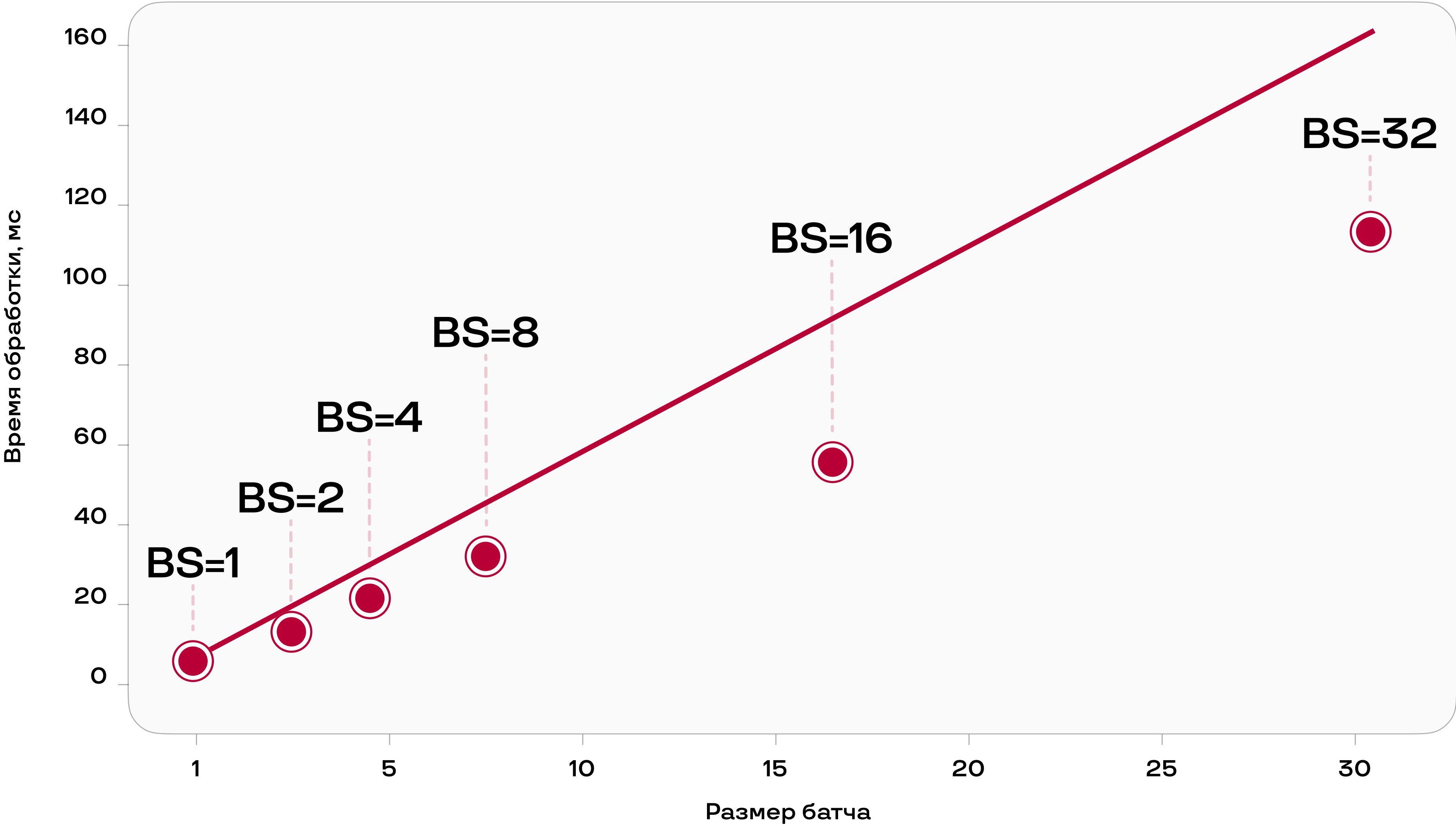


## На что обратить внимание:

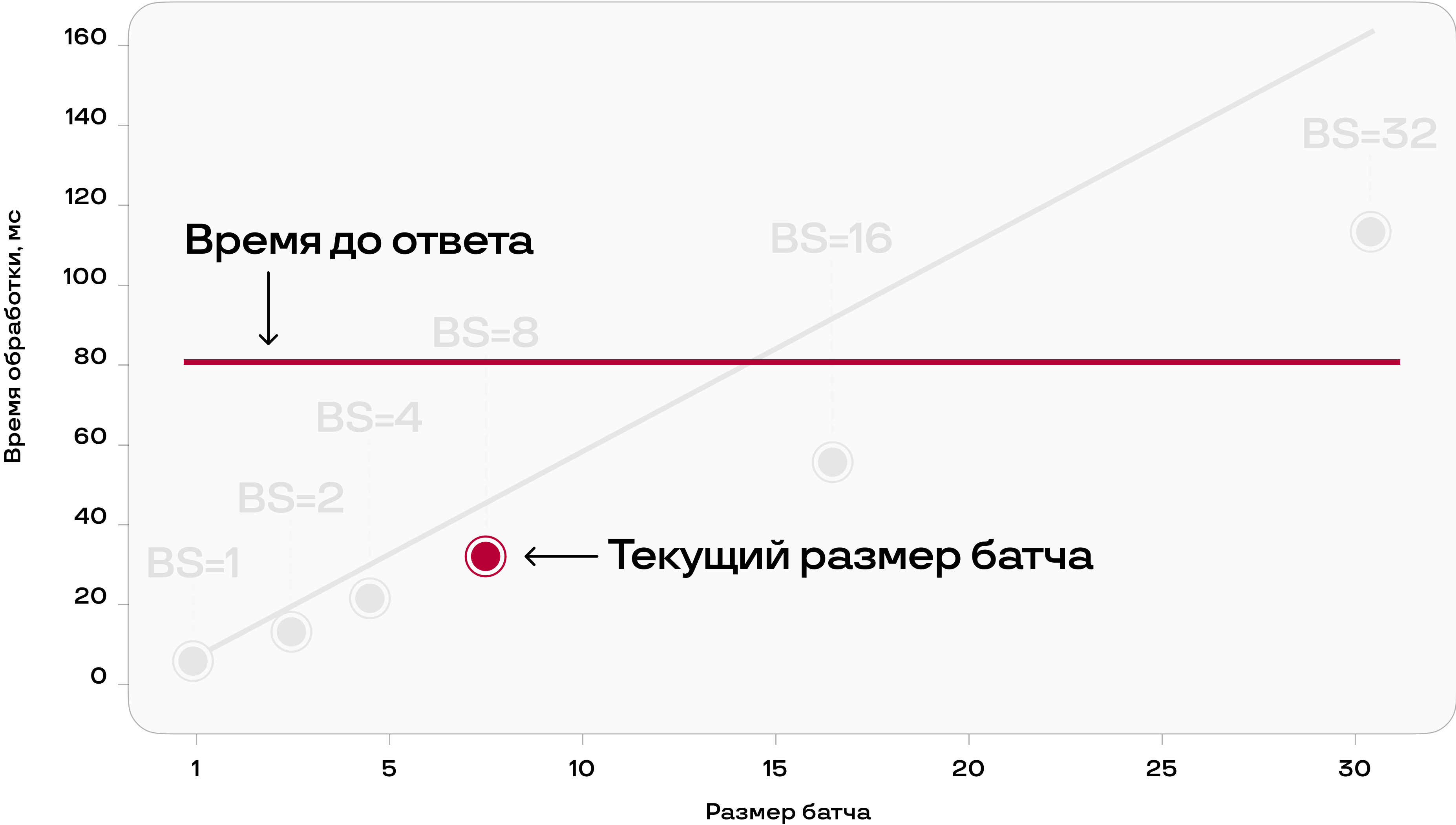
- 1 Оптимальные параметры
- 2 Динамический батч
- 3 Квантизация



# Оптимизация инференса: динамический батч

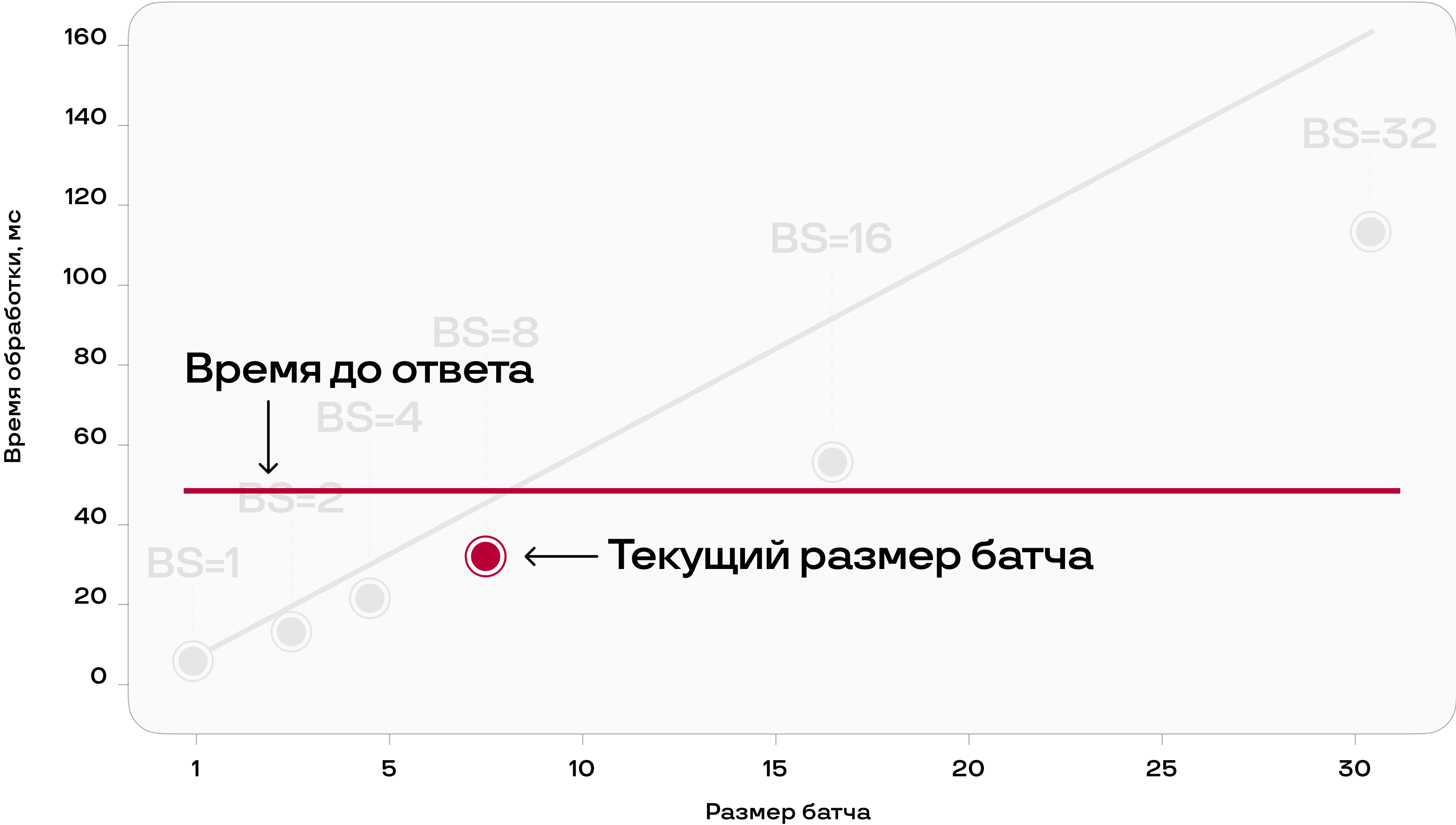


# Оптимизация инференса: динамический батч





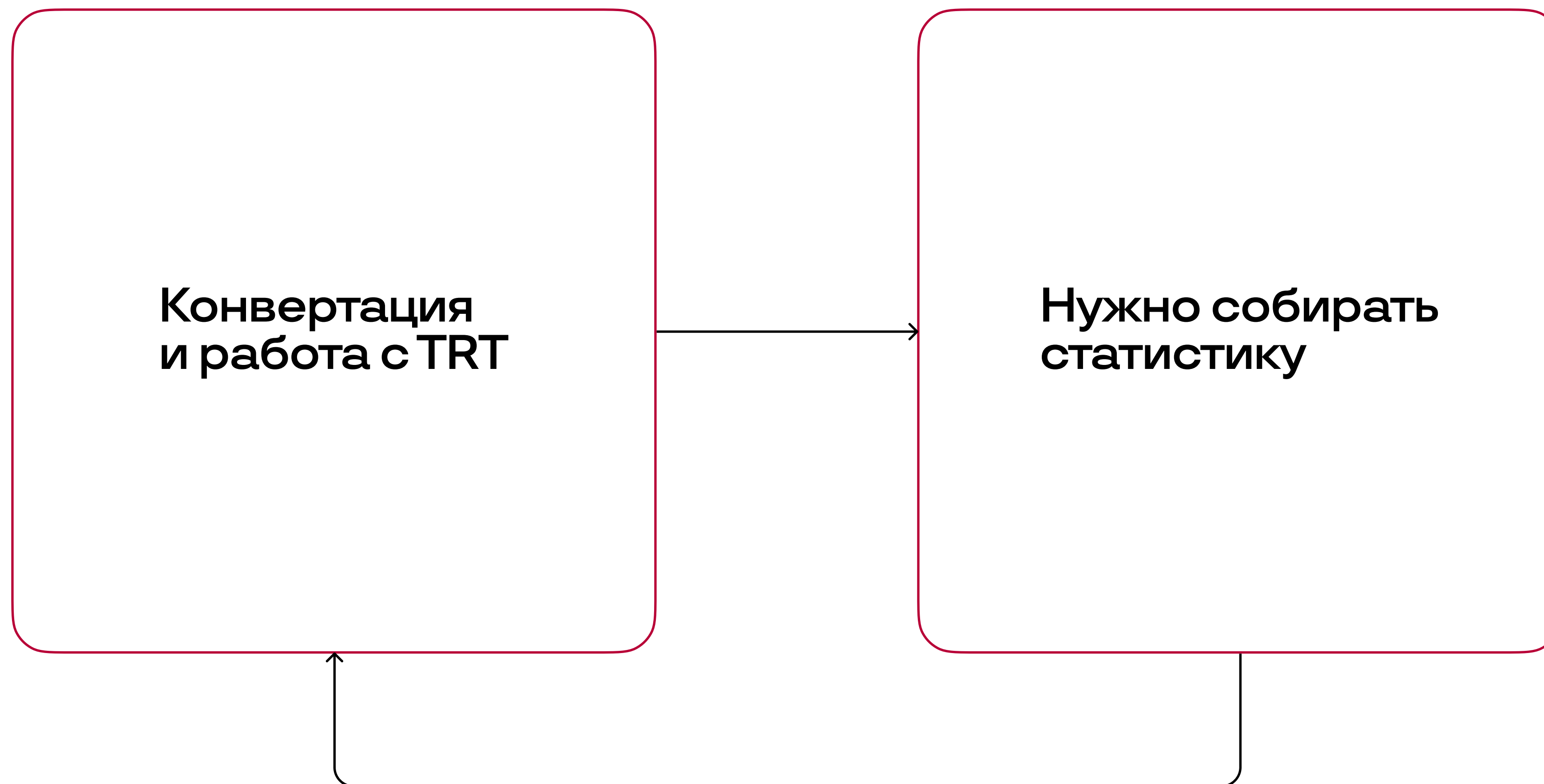
# Оптимизация инференса: динамический батч



# Батчинг



# Особенности динамического батчинга





# Варианты оптимизации

Модель

Инференс и батчинг

Вспомогательный код

# Варианты оптимизации

Модель

Инференс и батчинг

Вспомогательный код

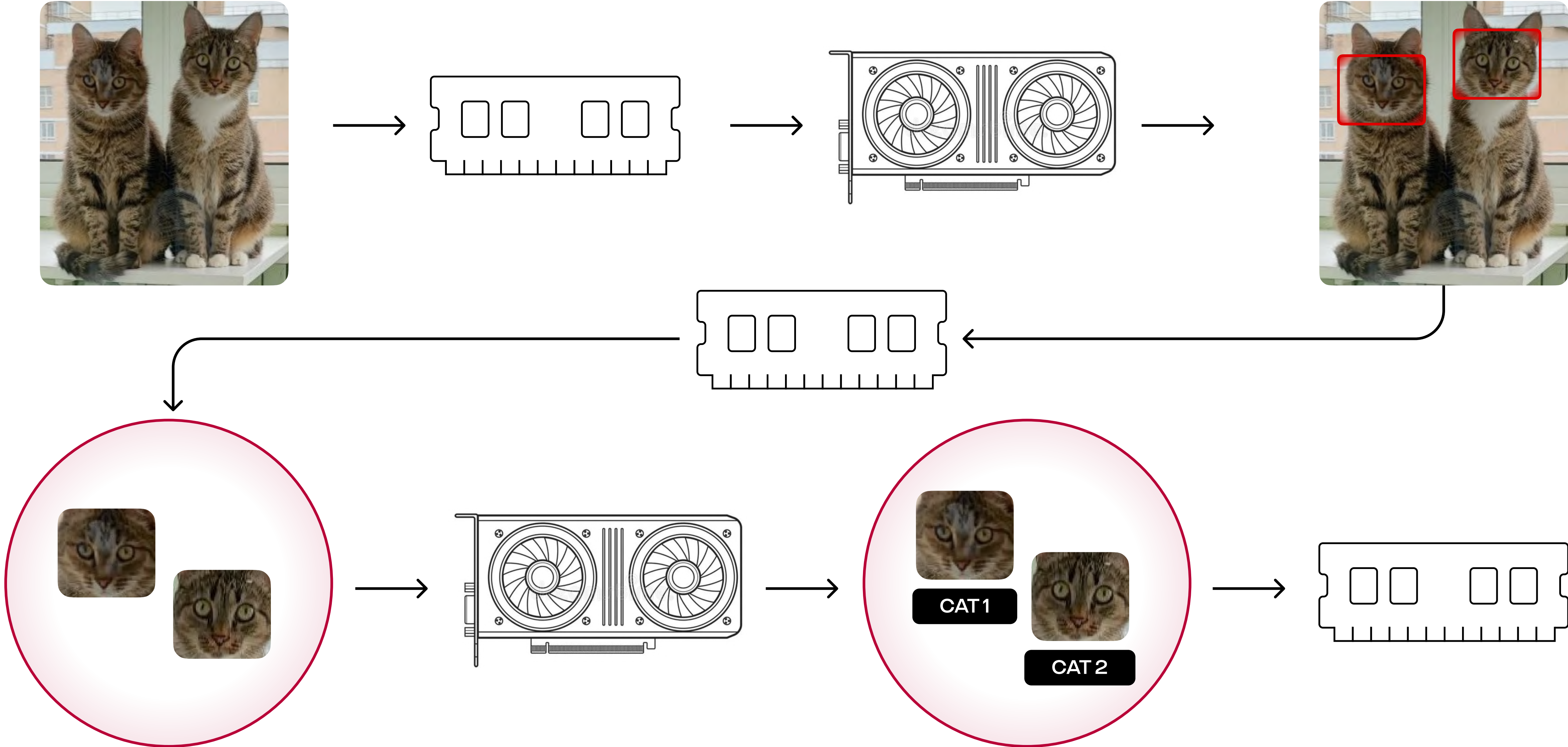
# Оптимизация кода: C++-модуль для Python

Все эти операции целиком выполняются на GPU:

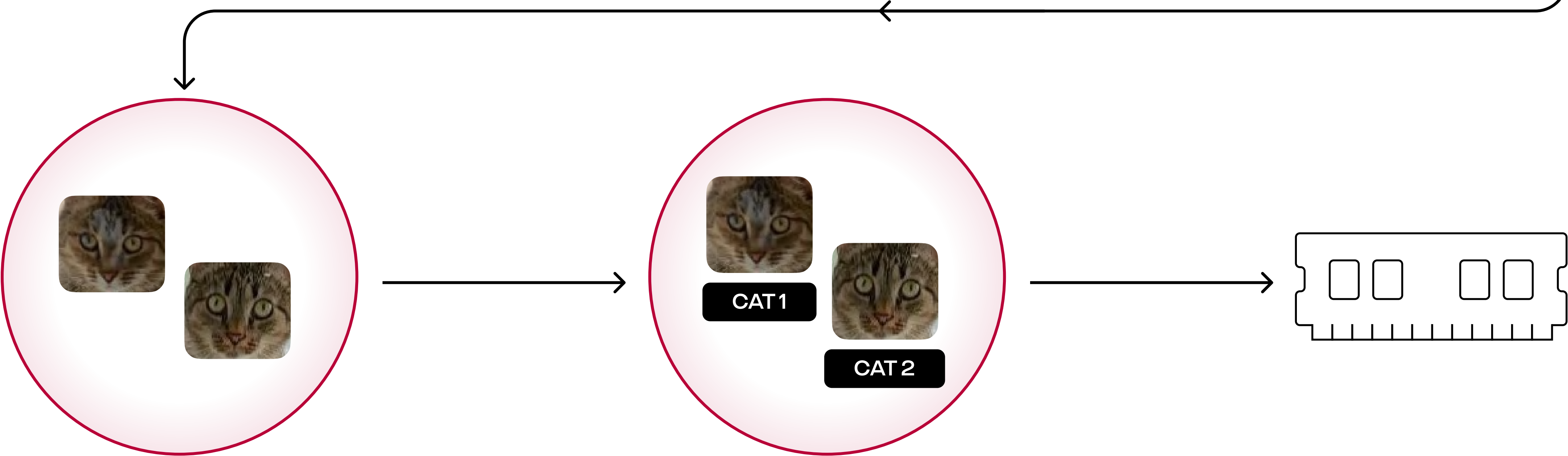
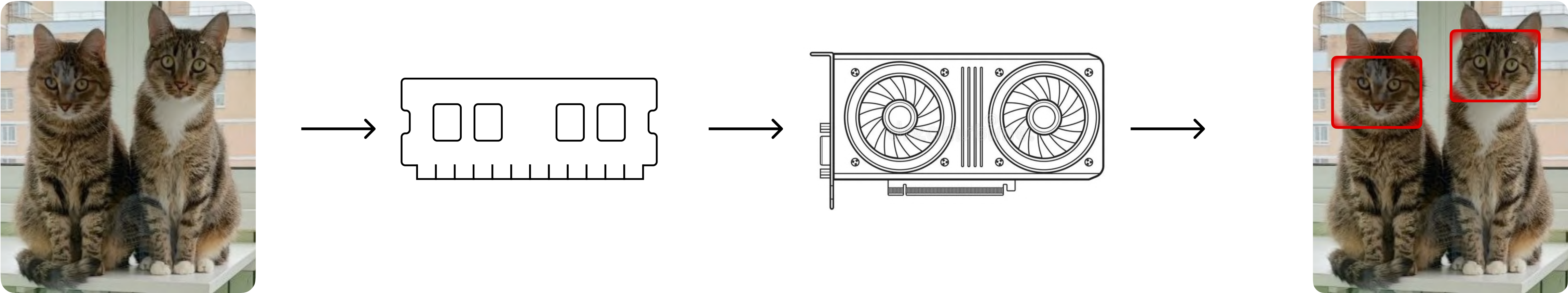
- 1 JPEG-кодирование / декодирование
- 2 Любая обработка изображений на OpenCV: обрезка, экстракция масок, нормализация, и т.д.
- 3 Конвертация: OpenCV Mat  $\leftrightarrow$  TRT / Torch Tensor
- 4 Инференс моделей на TensorRT
- 5 Центральный менеджер памяти GPU



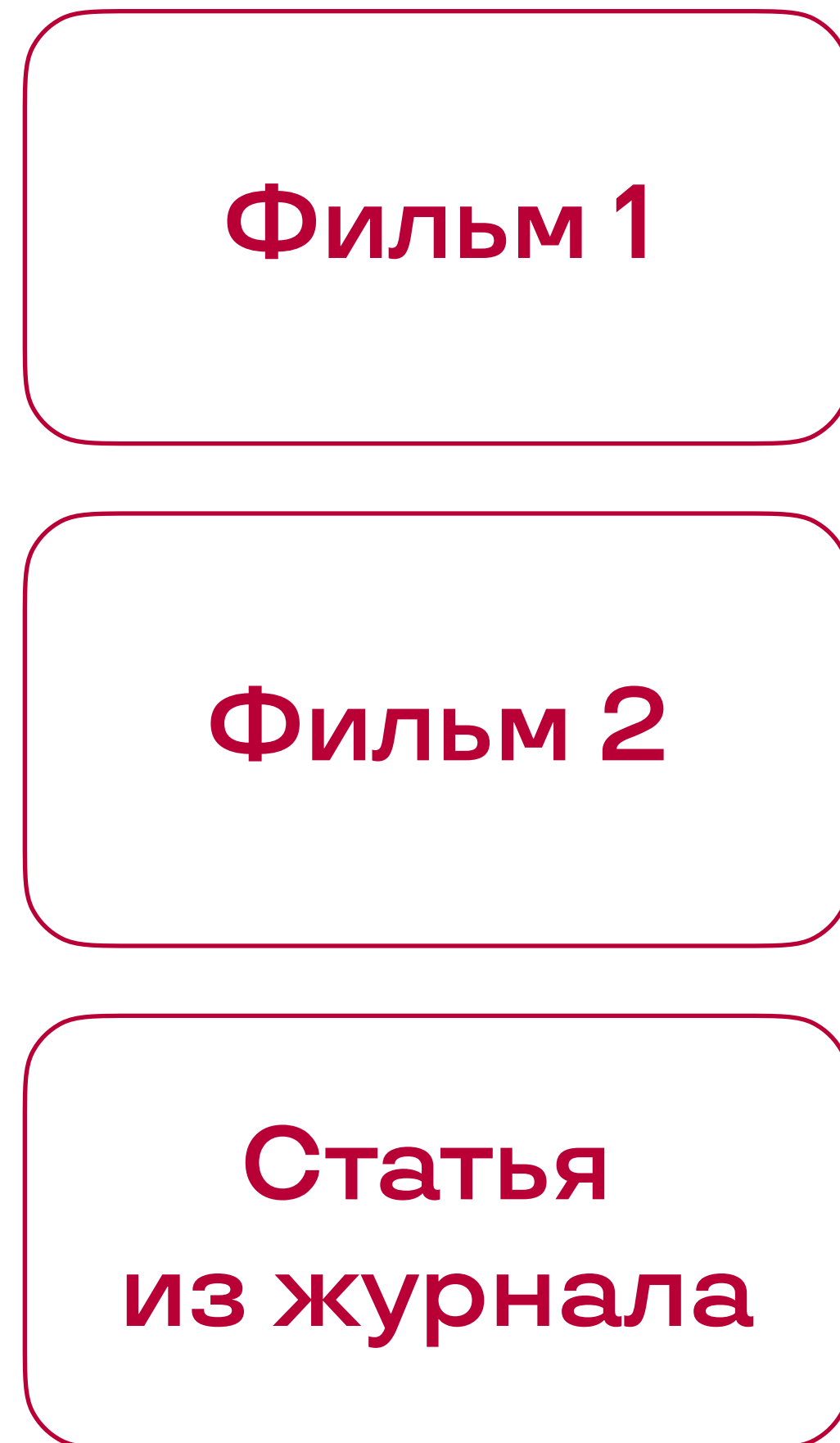
Почему это важно?



# Почему это важно?

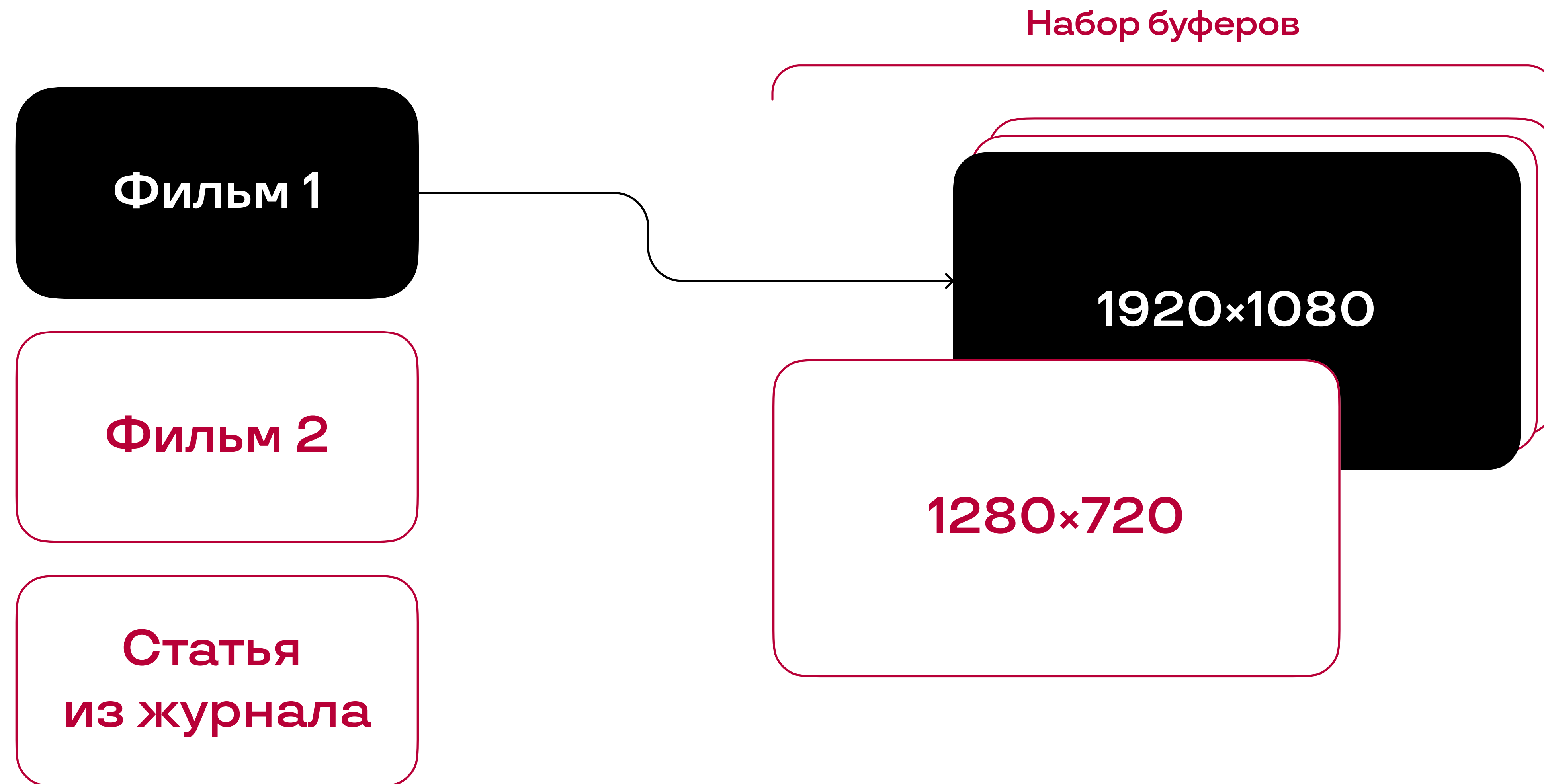


# Лишние аллокации

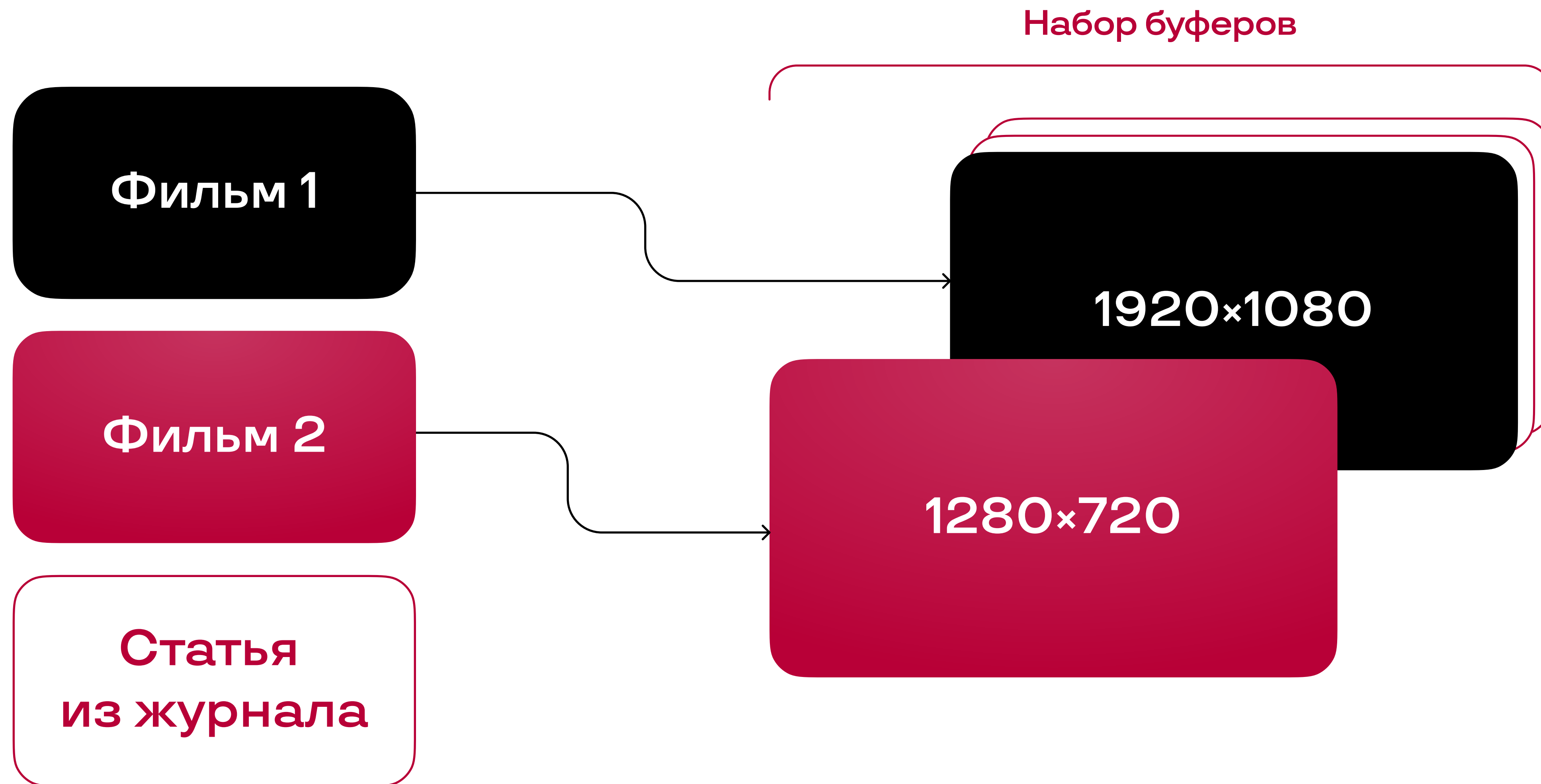




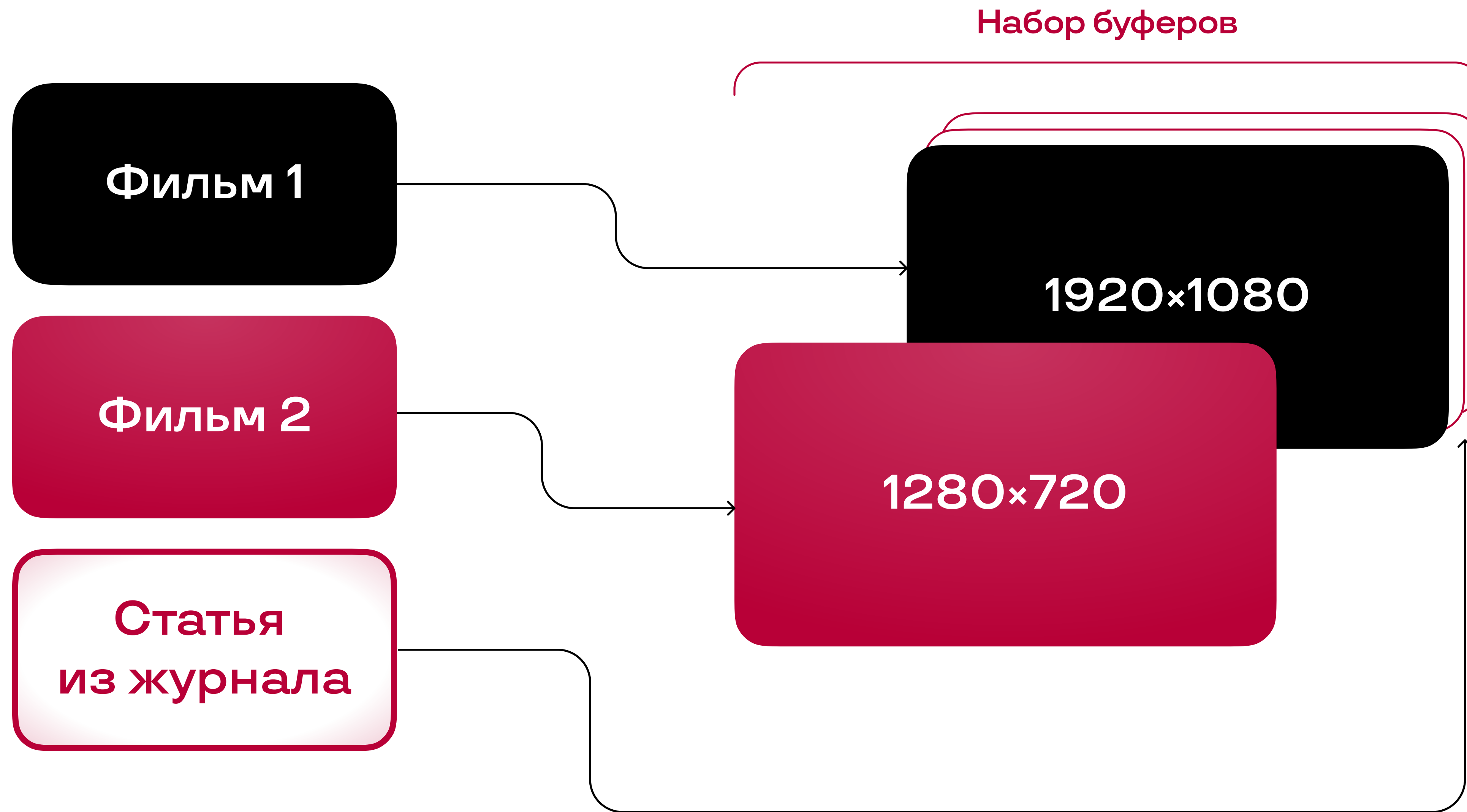
# Лишние аллокации



# Лишние аллокации



# Лишние аллокации



# Почему не Nvidia Triton Inference Server

38

**Triton Inference  
Server**

**VS**

**Своя обработка ошибок**

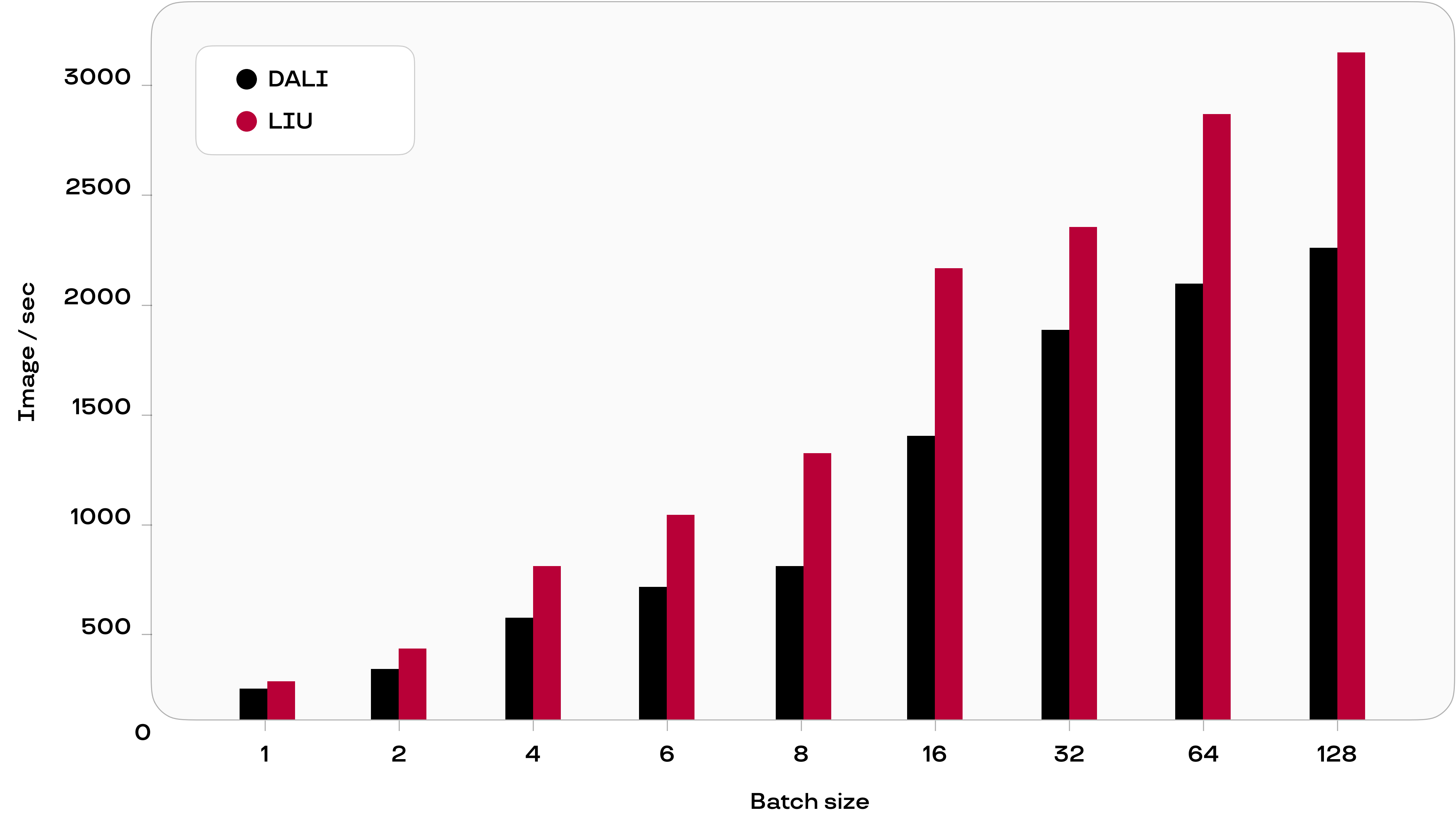
**Гибкость и свои интерфейсы**

**Легковесность**

**Скорость**



# Сравнение скорости декодирования

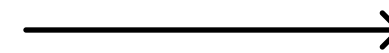


# Использование CPU

Перенос инференса и всех операций на GPU

# Использование CPU

Перенос инференса и всех операций на GPU



Большая загрузка карты при «простаивании» CPU

# Использование CPU

42

Перенос инференса и всех операций на GPU



Большая загрузка карты при «простаивании» CPU

Передача части операций CPU



# Использование CPU

Перенос инференса и всех операций на GPU



Большая загрузка карты при «простаивании» CPU

Передача части операций CPU



Более равномерная загрузка девайсов

# Использование CPU

44

Перенос инференса и всех операций на GPU



Большая загрузка карты при «простаивании» CPU

Передача части операций CPU



Более равномерная загрузка девайсов

Чем можно загрузить CPU?

- 1 Распаковка png и webp
- 2 Попытка восстановления битых jpeg
- 3 Гибридный режим для декодирования

# 1 Совсем другой язык: Julia, Golang, Rust, C++

**1** Совсем другой язык: Julia, Golang, Rust, C++

**2** Почти питон: Numba, Cython



- 1** Совсем другой язык: Julia, Golang, Rust, C++
- 2** Почти питон: Numba, Cython
- 3** Модули на C++, основной Python

# Варианты оптимизации

48

**Модель**

**Инференс и батчинг**

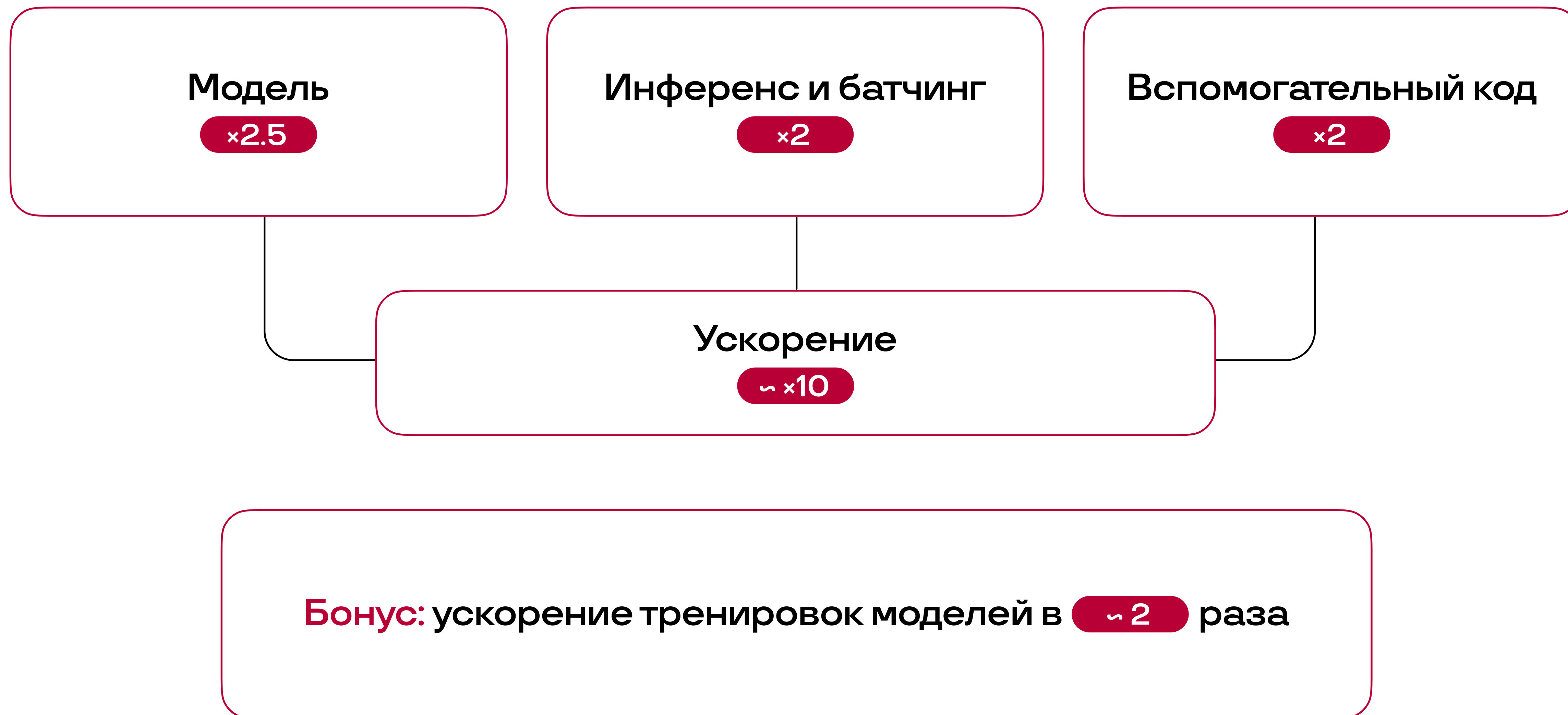
**Вспомогательный код**

# Варианты оптимизации

49



# Варианты оптимизации





Спасибо за внимание!

Оценить доклад

Telegram  
[@Grigoriy\\_Alekseenko](https://t.me/Grigoriy_Alekseenko)

Канал команды в Telegram  
[t.me/layercv](https://t.me/layercv)

